CLAIMS

What is claimed is:

1. A method to process a document, comprising:

partitioning document text and assigning semantic meaning to words, where assigning comprises applying a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments;

recognizing any substructures present in the chemical name fragments; and

determining structural connectivity information of the chemical name fragments and recognized substructures and storing the determined structural connectivity information in a searchable index.

2. A method as in claim 1, further comprising searching the index by at least one of fragment name and substructure name.

3. A method as in claim 1, further comprising searching the index by at least one of fragment connectivity and substructure connectivity.

4. A method as in claim 1, further comprising searching the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

5. A method as in claim 1, further comprising searching the index by at least one of fragment and substructure connectivity using a graphical user interface.

6. A method as in claim 1, where the determined structural connectivity information is stored in a searchable structure index, further comprising storing text associated with processed documents in a text index, and searching the text index using at least one of a fragment name and a substructure name and searching the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, identifying at least one document that contains a reference to a corresponding chemical compound.

7. A method as in claim 1, where determining structural connectivity information comprises looking up recognized fragments and substructures in a structure dictionary.

8. A method as in claim 7, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.

9. A method as in claim 1, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

10. A method as in claim 1, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

11. A method as in claim 1, further comprising filtering recognized chemical name fragments using

a list of stop words to eliminate erroneous chemical name fragments.

12. A method as in claim 1, where chemical name fragments are further recognized by using common chemical word endings.

13. A method as in claim 1, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

14. A method as in claim 1, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

15. A method as in claim 14, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

16. A method as in claim 14, where the characters comprise at least one of upper case C, O, R, N and H.

17. A method as in claim 14, where the characters comprise strings of at least one of lower case xy, ene, ine, yl, ane and oic.

18. A method as in claim 1, comprising an initial step of tokenizing the document to provide a sequence of tokens.

19. A system to process a document, comprising:

a unit to partition document text and to assign semantic meaning to words, where assigning comprises applying a plurality of regular expressions, rules and a plurality of dictionaries to recognize chemical name fragments;

a unit to recognize any substructures present in the chemical name fragments; and

a unit to determine structural connectivity information of the chemical name fragments and recognized substructures and to store the determined structural connectivity information in a searchable index.

20. A system as in claim 19, further comprising searching the index by at least one of fragment name and substructure name.

21. A system as in claim 19, further comprising unit to search the index by at least one of fragment connectivity and substructure connectivity.

22. A system as in claim 19, further comprising a unit to search the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

23. A system as in claim 19, further comprising a unit to search the index by at least one of fragment

and substructure connectivity using a graphical user interface.

24. A system as in claim 19, where the determined structural connectivity information is stored in a searchable structure index, further comprising a unit to store text associated with processed documents in a text index, and a unit to search the text index using at least one of a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

25. A system as in claim 19, where said unit that determines structural connectivity information looks up recognized fragments and substructures in a structure dictionary.

26. A system as in claim 25, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.

27. A system as in claim 19, where said plurality of dictionaries comprise a dictionary of common chemical prefixes and a dictionary of common chemical suffixes.

28. A system as in claim 19, where said plurality of dictionaries comprise a dictionary of stop words to eliminate erroneous chemical name fragments.

29. A system as in claim 19, further comprising a unit to filter recognized chemical name fragments

using a list of stop words to eliminate erroneous chemical name fragments.

30. A system as in claim 19, where chemical name fragments are further recognized by using common chemical word endings.

31. A system as in claim 19, where application of said regular expressions and rules results in punctuation characters being one of maintained or removed between chemical name fragments as a function of context.

32. A system as in claim 19, where said regular expressions comprise a plurality of patterns, individual ones of which are comprised of at least one of characters, numbers and punctuation.

33. A system as in claim 32, where the punctuation comprises at least one of parenthesis, square bracket, hyphen, colon and semi-colon.

34. A system as in claim 32, where the characters comprise at least one of upper case C, O, R, N and H.

35. A system as in claim 32, where the characters comprise strings of at least one of lower case xy, ene, ine, yl, ane and oic.

36. A system as in claim 19, further comprising an input tokenizer unit to receive documents to be processed to provide a sequence of tokens.

37. A computer program product for storing in a computer readable form a set of computer program instructions for directing at least one computer to process a text document, comprising instructions to parse document text to recognize chemical name fragments; instructions to recognize any substructures present in the chemical name fragments; and instructions to determine structural connectivity information of the chemical name fragments and recognized substructures and to store the determined structural connectivity information in a searchable index.

38. A computer program product as in claim 37, further comprising instructions to search the index by at least one of fragment name and substructure name.

39. A computer program product as in claim 37, further comprising instructions to search the index by at least one of fragment connectivity and substructure connectivity.

40. A computer program product as in claim 37, further comprising instructions to search the index by a combination of at least one of fragment and substructure name, and at least one of fragment and substructure connectivity.

41. A computer program product as in claim 37, further comprising instructions to search the index by at least one of fragment and substructure connectivity using a graphical user interface.

42. A computer program product as in claim 37, where the determined structural connectivity information is stored in a searchable structure index, further comprising instructions to store text associated with processed documents in a text index, and instructions to search the text index using

at least one of a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

43. A system comprising a plurality of computers at least two of which are coupled together through a data communications network, said system comprising a unit to parse document text recognize chemical name fragments; a unit to recognize any substructures present in the chemical name fragments; and a unit to determine structural connectivity information of the chemical name fragments and recognized substructures and to store the determined structural connectivity information in a searchable index.

44. A system as in claim 43, where the determined structural connectivity information is stored in a searchable structure index, further comprising a unit to store text associated with processed documents in a text index, and a unit to search the text index using at least one of a fragment name and a substructure name and to search the structure index by at least one of fragment connectivity and substructure connectivity, and at an intersection of the search results from the structure index and the text index, to identify at least one document that contains a reference to a corresponding chemical compound.

45. A system as in claim 43, where said unit that determines structural connectivity information looks up recognized fragments and substructures in a structure dictionary.

46. A system as in claim 45, where the structure dictionary comprises at least one of a MOL dictionary and a SMILES dictionary.